Projection-based Regularized Dual Averaging for Stochastic Optimization with Its Applications to Classification and Regression

Asahi Ushio

ID: 81615354
Masahiro Yukawa LAB
Dept. Electronics and Electrical Engineering, Keio University

Final Presentation for Master Study
December 23, 2017

Table of Contents: Master Thesis

Focus in This Talk

Introduction

Background Era of data deluge (SNS, IoT, Digital News) [1].
$\rightarrow$ **Machine learning and signal processing become more important !**

Challenges
- **Real-time streaming data.**
- **High-dimensional data.**
- **Low complexity desired.**

Regularized Stochastic Optimization
- **Online learning.**
- **Reduce the estimation variance.**
- **Sparse solution.**

[1] McKinsey, "Big Data: The next frontier for innovation, competition, and productivity," 2011.

In Signal Processing ...
- **Squared distance cost:** Projection-based method NLMS, APA, APFBS [2]
- **Change Geometry:** PAPA, Variable Metric [3]

In Machine Learning ...
- **Regularized Dual Averaging (RDA) type:** RDA [4]
- **Forward Backward Splitting (FBS) type:** FOBOS [5]
- **Change Geometry:** AdaGrad [6]

[2] Murakami *et al.*, 2010, [3] Yukawa *et al.*, 2009, [4] Xiao, 2010, [5] Singer *et al.*, 2009, [6] Duchi *et al.*, 2011

In Machine Learning ...

Background Now, we have many kinds of data (SNS, IoT, Digital News) [1].
$\rightarrow$ **Machine learning and signal processing become more important !**

Challenges
- **Real-time streaming data**.
- **High-dimensional data**.
- **Low complexity desired**.

Regularized Stochastic Optimization
- **Online learning**.
- **Reduce the estimation variance**.
- **Sparse solution**.

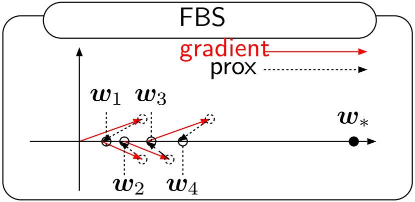
[1] McKinsey, "Big Data: The next frontier for innovation, competition, and productivity," 2011.

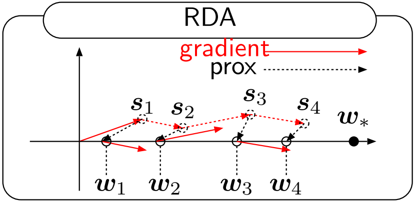
In Signal Processing ...
- **Squared distance cost:** Projection-based method NLMS, APA, APFBS [2]
- **Change Geometry:** PAPA, Variable Metric [3]

In Machine Learning ...
- **Regularized Dual Averaging (RDA) type:** RDA [4]
- **Forward Backward Splitting (FBS) type:** FOBOS [5]
- **Change Geometry:**AdaGrad [6]

[2] Murakami *et al.*, 2010, [3] Yukawa *et al.*, 2009, [4] Xiao, 2010, [5] Singer *et al.*, 2009, [6] Duchi *et al.*, 2011

FBS type versus RDA type





Accumulation of prox

FBS : Accumulation of prox
$\Rightarrow$ Increase bias
$\Rightarrow$ Low accuracy

RDA : Accumulation **free**
$\Rightarrow$ High accuracy

$\Downarrow$

	Sparsity	Accuracy
FBS	○	△
RDA	○	◎

Contribution of My Work

Motivation

1. **RDA** $\Rightarrow$ Sparse solution.
2. **Squared distance cost** $\Rightarrow$ Stability.
3. **Variable metric** $\Rightarrow$ Acceleration of the convergence by changing the geometry.

Integrate the three ideas

$\longrightarrow$ **Sparsity-promoting and stable learning !**

Proposed Algorithm

Projection-based Regularized Dual Averaging (PDA)

- Features: **RDA** with **Squared distance cost**, and **variable metric**.
- Show efficacy by extensive simulations (classification and regression).

Problem Setting

- **Model:** For input $\boldsymbol{x}_t \in \mathbb{R}^n$ at time t and coefficient $\boldsymbol{w} \in \mathbb{R}^n$,

$$\hat{y}_t = \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_t, \tag{1}$$

where $\hat{y}_t$ is an estimation of output (label) y_t.

- **Loss:**

$$l_t(\boldsymbol{w}) = \underbrace{\varphi_t(\boldsymbol{w})}_{\text{loss}} + \underbrace{\psi_t(\boldsymbol{w})}_{\text{regularizer}} . \tag{2}$$

1. **Ordinary loss function**

$$\text{Regression} : \varphi_t(\boldsymbol{w}) = (y_t - \hat{y}_t)^2/2.$$

$$\text{Classification} : \varphi_t(\boldsymbol{w}) = y_t \log\left(1 + e^{-\hat{y}_t}\right) + (1 - y_t)\log\left(\frac{1 + e^{-\hat{y}_t}}{e^{-\hat{y}_t}}\right).$$

2. **Projection**

$$\text{Regression} : \varphi_t(\boldsymbol{w}) = \frac{(y_t - \hat{y}_t)^2}{2\,||\boldsymbol{x}_t||^2}.$$

$$\text{Classification} : \varphi_t(\boldsymbol{w}) = \frac{\left([\hat{y}_t y_t - 1]_+\right)^2}{2\,||\boldsymbol{x}_t||^2}.$$

RDA (Regularized Dual Averaging) [1]

Framework of RDA

$$\boldsymbol{w}_t := \arg\min_{\boldsymbol{w}\in\mathbb{R}^n} \left(\left\langle \frac{\sum_{i=1}^t \boldsymbol{g}_i}{t}, \boldsymbol{w} \right\rangle + \frac{\beta_t}{t} h(\boldsymbol{w}) + \psi(\boldsymbol{w}) \right) \qquad (3)$$

$h(\boldsymbol{w})$: distance function, $(\beta_\tau)_{\tau=1,\dots,t}$: nonnegative nondecreasing sequence.

- In the case of $h(\boldsymbol{w}) = ||\boldsymbol{w}||^2/2$, with $\boldsymbol{s}_t = \sum_{i=1}^t \boldsymbol{g}_i$,

$$\boldsymbol{w}_t := \arg\min_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{1}{2} \left|\left| \boldsymbol{w} + \frac{1}{\beta_t}\boldsymbol{s}_t \right|\right|^2 + \frac{t}{\beta_t}\psi(\boldsymbol{w}) \right) = \mathrm{prox}_{\frac{t}{\beta_t}\psi}\left(-\frac{1}{\beta_t}\boldsymbol{s}_t \right). \qquad (4)$$

Definition: Proximity operator of ψ of index $\eta > 0$

$$\mathrm{prox}_{\eta\psi}(\boldsymbol{w}) := \arg\min_{\boldsymbol{z}\in\mathbb{R}^n} \left(\eta\psi(\boldsymbol{w}) + \tfrac{1}{2}||\boldsymbol{w}-\boldsymbol{z}||^2 \right), \quad \forall \boldsymbol{w} \in \mathbb{R}^n.$$

[1] Xiao "Dual averaging methods for regularized stochastic learning and online optimization." Journal of Machine Learning Research 2010

RDA (Regularized Dual Averaging) [1]

Framework of RDA

$$\boldsymbol{w}_t := \underset{\boldsymbol{w}\in\mathbb{R}^n}{\arg\min} \left(\left\langle \frac{\sum_{i=1}^t \boldsymbol{g}_i}{t}, \boldsymbol{w} \right\rangle + \frac{\beta_t}{t} h(\boldsymbol{w}) + \psi(\boldsymbol{w}) \right) \tag{5}$$

$h(\boldsymbol{w})$: distance function, $(\beta_\tau)_{\tau=1,\dots,t}$: nonnegative nondecreasing sequence.

In the case of $h(\boldsymbol{w}) = ||\boldsymbol{w}||^2/2$, with $\boldsymbol{s}_t = \sum_{i=1}^t \boldsymbol{g}_i$,

$$\boldsymbol{w}_t := \underset{\boldsymbol{w}\in\mathbb{R}^n}{\arg\min} \left(\frac{1}{2} \left\| \boldsymbol{w} + \frac{1}{\beta_t}\boldsymbol{s}_t \right\|^2 + \frac{t}{\beta_t}\psi(\boldsymbol{w}) \right) = \mathrm{prox}_{\frac{t}{\beta_t}\psi}\left(-\boxed{\frac{1}{\beta_t}}\boldsymbol{s}_t \right).$$

- $\beta_t \sim \mathcal{O}(\sqrt{t}) \implies 1/\beta_t \sim \mathcal{O}(1/\sqrt{t})$, regularizer $\sim \mathcal{O}(\sqrt{t})$.
- If β_t stays constant $\implies 1/\beta_t =$ constant, regularizer $\sim \mathcal{O}(t) \Rightarrow$ Too sparse !

Proposed Algorithm

Conventional RDA $w_t := \arg\min_{w \in \mathbb{R}^n} \left(\frac{1}{2} \left\| w + \boxed{\frac{1}{\beta_t}} s_t \right\|^2 + \boxed{\frac{t}{\beta_t}} \psi(w) \right).$

Projection-based Dual Averaging (PDA)

$$w_t := \arg\min_{w \in \mathbb{R}^n} \left(\frac{1}{2} \left\| w + \boxed{\eta} s_t \right\|_{Q_t}^2 + \boxed{\eta} \psi_t(w) \right) = \mathrm{prox}_{\eta\psi_t}^{Q_t}(-\eta s_t)$$

"RDA + Projection + sparsity-promoting metric"

- Constant reg. parameter and constant step size. $\Rightarrow$ Sparse solution with high accuracy.
- **Projection (Note that g_t in RDA is a subgradient)**:
  - Variable metric: Sparsity enhancement.
  - Squared distance cost: Robustness to input fluctuation and noise.

$Q_t \in \mathbb{R}^{n \times n}$: A positive definite matrix.
$\|w\|_{Q_t} := \sqrt{\langle w, w \rangle_{Q_t}}$, $\langle w, z \rangle_{Q_t} := w^\top Q_t z$ for $w, z \in \mathbb{R}^n$.
$\mathrm{prox}_{\eta\psi_t}^{Q_n}(w) := \arg\min_{z \in \mathbb{R}^n} \left(\eta\psi_t(z) + \frac{1}{2} \|w - z\|_{Q_n}^2 \right), \quad w \in \mathbb{R}^n.$

Relation to Prior Work

SGD type: SGD, **NLMS**, **APA**, **PAPA**, Adam, AdaDelta

Sparsity-promoting: FBS (Forward Backward Splitting) type

FOBOS [1], AdaGrad-FBS [2], **APFBS** [3]

Dual Averaging type: Dual Averaging [4]

Sparsity-promoting: RDA (Regularized Dual Averaging) type

RDA [5], AdaGrad-RDA [2]

*bold : projection-based method

[1] Singer *et al.*, 2009 [2] Duchi *et al.*, 2011 [3] Murakami *et al.*, 2010 [4] Nesterov, 2009 [5] Xiao, 2009

	Ordinary loss	Projection-based
FBS type	FOBOS, AdaGrad-FBS	**APFBS**
RDA type	RDA, AdaGrad-RDA	

Relation to Prior Work

SGD type: SGD, **NLMS**, **APA**, **PAPA**, Adam, AdaDelta

Sparsity-promoting: FBS (Forward Backward Splitting) type

FOBOS [1], AdaGrad-FBS [2], **APFBS** [3]

Dual Averaging type: Dual Averaging [4]

Sparsity-promoting: RDA (Regularized Dual Averaging) type

RDA [5], AdaGrad-RDA [2], **PDA**

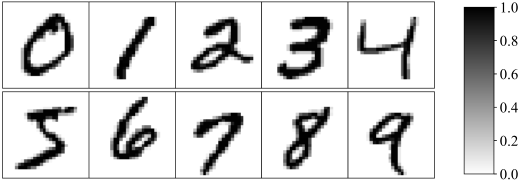
*bold : projection-based method

[1] Singer *et al.*, 2009 [2] Duchi *et al.*, 2011 [3] Murakami *et al.*, 2010 [4] Nesterov, 2009 [5] Xiao, 2009

	Ordinary Cost Function	Projection-based
FBS type	FOBOS, AdaGrad-FBS	**APFBS**
RDA type	RDA, AdaGrad-RDA	**PDA**

Experiment: Classifications

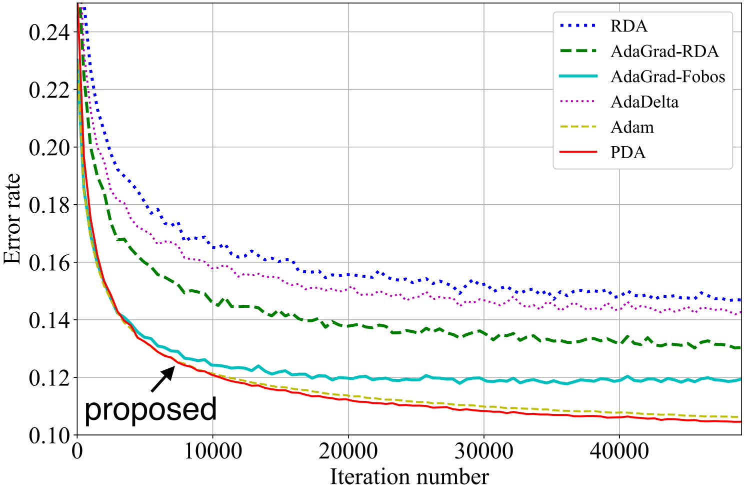
- **Hand written digit (MNIST)**



- **News text (RCV)**
  - Label: Economics, Industrial, Markets, and Social (multi label).
  - Data: News text (800,000 records).

Experiment: Hand Written Digit Classification (MNIST)

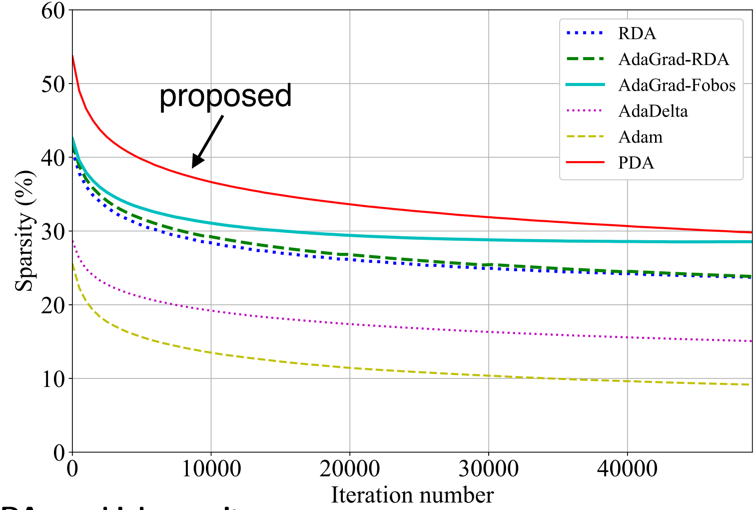
Error Rate



- **PDA shows the best performance.**

Experiment: Hand Written Digit Classification (MNIST)

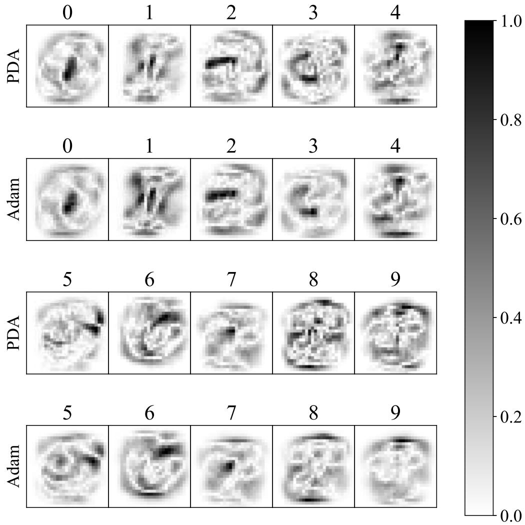
Proportion of the Zero Components of the Estimated Coefficient Vector



- **PDA $\longrightarrow$ high sparsity.**

Experiment: Hand Written Digit Classification (MNIST)

Visualization of the Estimated Coefficient for Adam and PDA

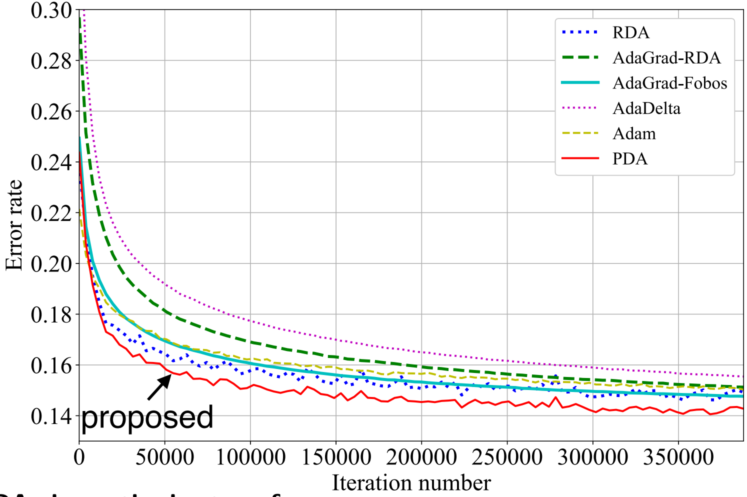


- Magnitude of each coefficient.
- Normalized in $[0, 1]$.

- PDA's solution focuses on **the edge and hole** of the digits.

Experiment: News Text Classification (RCV)

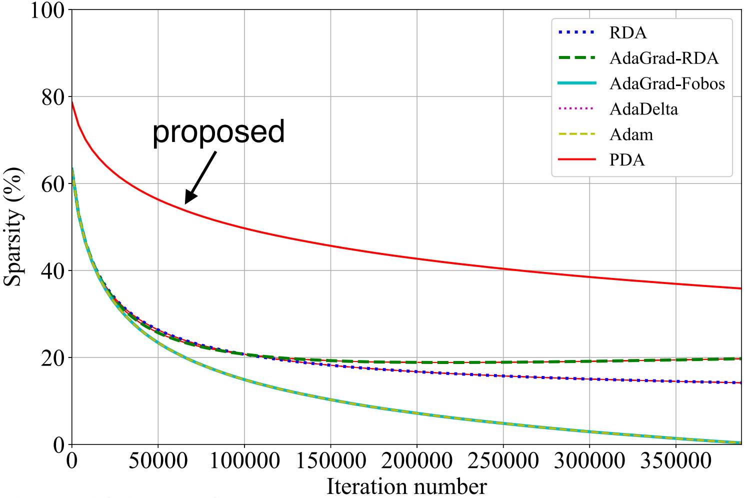
Error Rate



- **PDA shows the best performance.**

Experiment: News Text Classification (RCV)

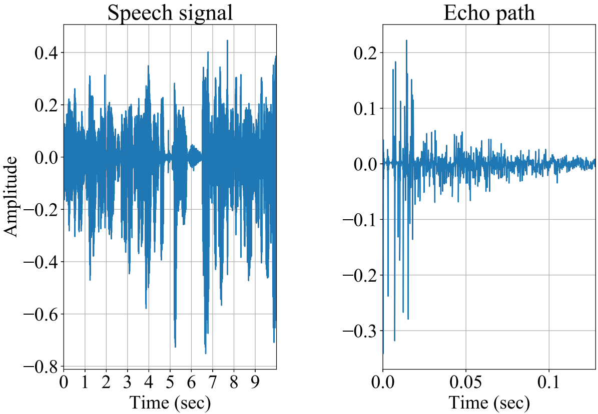
Proportion of the Zero Components of the Estimated Coefficient Vector



- **PDA $\longrightarrow$ high sparsity.**

Experiment: Regressions

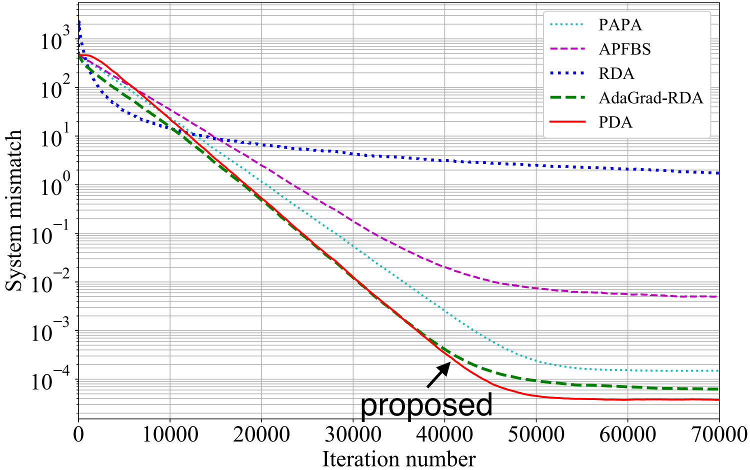
- **Sparse system estimation**
  - system: 1000 order, 80% coefficient zero.
- **Echo canceling**
  - echo path: 1024 order, frequency: 8kHz.



- **Nonlinear regression**

Experiment: Sparse -System Estimation

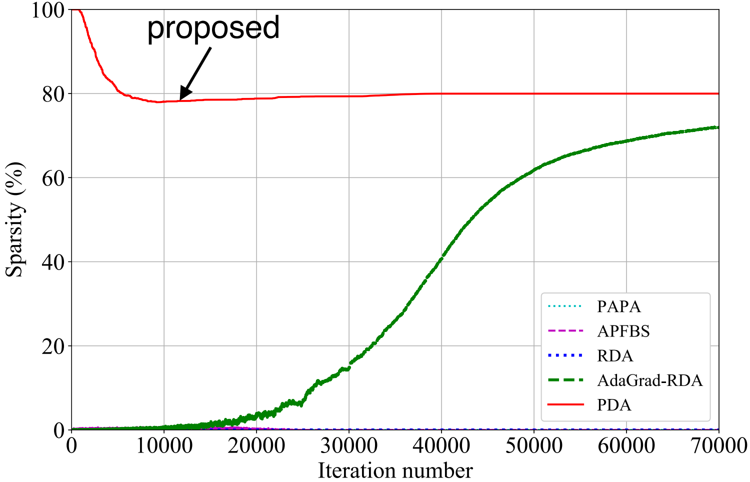
System Mismatch



- **PDA shows the best performance.**

Experiment: Sparse -System Estimation

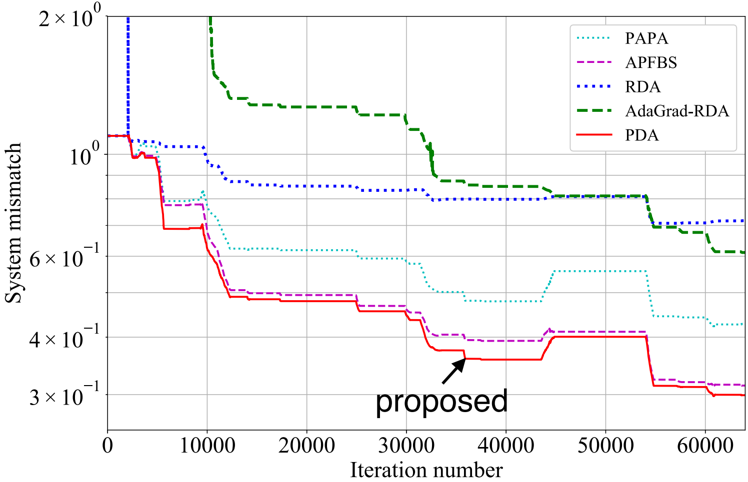
Proportion of the Zero Components of the Estimated Coefficient Vector



- **PDA achieves accurate sparsity.**

Experiment: Echo Cancellation

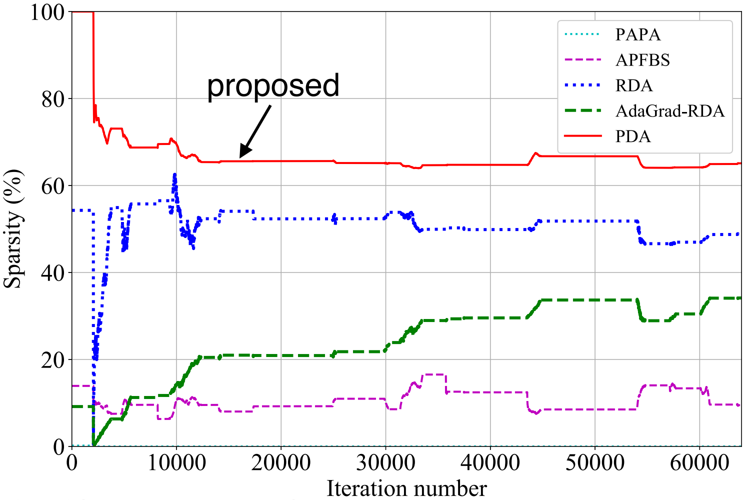
System Mismatch



- **PDA shows the best performance.**

Experiment: Echo Cancellation

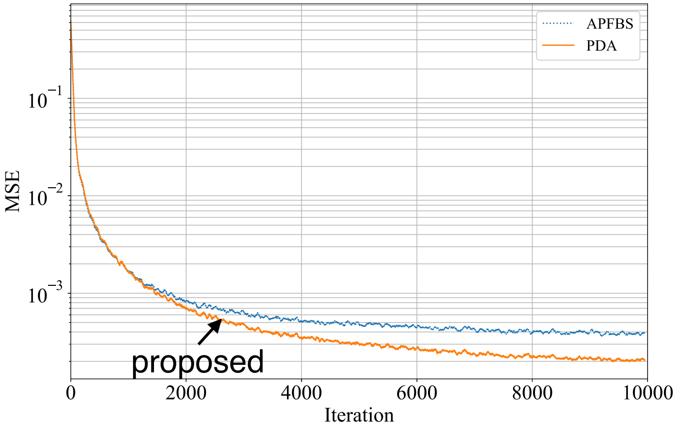
Proportion of the Zero Components of the Estimated Coefficient Vector



- **PDA achieves sparse solution.**

Experiment: Nonlinear Regression

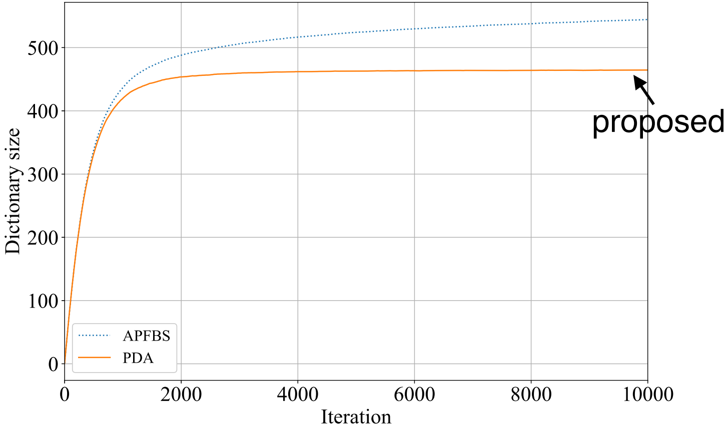
MSE



- **PDA shows better performance.**

Experiment: Nonlinear Regression

Dictionary Size



- **PDA achieves sparse solution.**

Conclusion

Conclusion

- Proposed the **projection-based regularized dual averaging (PDA)** algorithm.
  - **projection-based:** Input-vector normalization, stable adaptation by constant learning rate, the sparsity-seeking variable-metric.
  - **RDA:** Better sparsity-seeking.
- Various experiments demonstrated the efficacy of PDA.
  - **Online Classification:** MNIST (image recognition), RCV (text classification).
  - **Online regression:** Sparse system, Echo cancelling, nonlinear regression.

Qualifications

- **English Skill:** Toefl 88
- **Conference:** ICASSP 2017 oral presentation